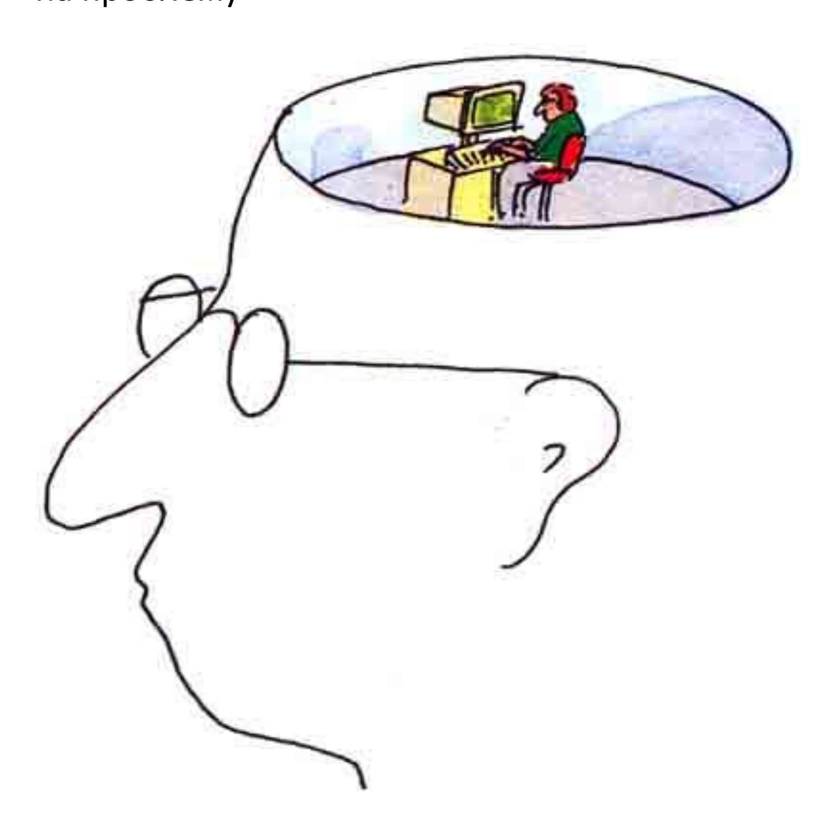
В МИРЕ НАУКИ. (Scientific American. Издание на русском языке). 1990. № 3

Искусственный интеллект: различные взгляды на проблему



Последние 35 лет попыток создать думающие машины были полны и удач, и разочарований. «Интеллектуальный» уровень современных компьютеров довольно высок, однако для того, чтобы компьютеры могли «разумно» вести себя в реальном мире, их поведенческие способности не должны уступать способностям по крайней мере самых примитивных животных. Некоторые специалисты, работающие в областях, не связанных с искусственным интеллектом, говорят, что компьютеры по своей природе не способны к сознательной умственной деятельности.

В этом номере журнала в статье Дж.Р. Сирла утверждается, что компьютерные программы никогда не смогут достичь разума в привычном для нас понимании. В то же время в другой статье, написанной П.М. Черчлендом и П. С. Черчленд, приводится мнение, что с помощью электронных схем, построенных по образу и подобию мозговых структур, возможно, удастся создать искусственный интеллект. За этим спором по существу скрывается вопрос о том, что такое мышление. Этот вопрос занимал умы людей на протяжении тысячелетий. Практическая работа с компьютерами, которые пока не могут мыслить, породила новый взгляд на этот вопрос и отвергла многие потенциальные ответы на него. Остается найти правильный ответ.

Разум мозга — компьютерная программа?

Нет. Программа лишь манипулирует символами, мозг же придает им смысл

ДЖОН СИРЛ

СПОСОБНА ли машина мыслить? Может ли машина иметь сознанные мысли в таком же смысле, в каком имеем их мы? Если под машиной понимать физическую систему, способную выполнять определенные функции (а что еще под ней можно понимать?), тогда люди — это машины особой, биологической разновидности, а люди могут мыслить, и, стало быть, машины, конечно, тоже могут мыслить. Тогда, по всей видимости, можно создавать мыслящие машины из самых разнообразных материалов — скажем, из кремниевых кристаллов или электронных ламп. Если это и окажется невозможным, то пока мы, конечно, этого еще не знаем.

Однако в последние десятилетия вопрос о том, может ли машина мыслить, приобрел совершенно другую интерпретацию. Он был подменен вопросом: способна ли машина мыслить только за счет выполнения заложенной в нее компьютерной программы? Является ли программа основой мышления? Это принципиально иной вопрос, потому что он не затрагивает физических, каузальных (причинных) свойств существующих или возможных физических систем, а скорее относится к абстрактным, вычислительным свойствам формализованных компьютерных программ, которые могут быть реализованы в любом материале, лишь бы он был способен выполнять эти программы.

Довольно большое число специалистов по искусственному интеллекту (ИИ) полагают, что на второй вопрос следует ответить положительно; другими словами, они считают, что составив правильные программы с правильными входами и выходами, они действительно создадут разум. Более того, они полагают, что имеют в своем распоряжении научный тест, с помощью которого можно судить об успехе или неудаче такой попытки. Имеется в виду тест Тьюринга, изобретенный Аланом М. Тьюрингом, основоположником искусственного интеллекта. Тест Тьюринга в том смысле, как его сейчас понимают, заключается просто в следующем: если компьютер способен демонстрировать поведение, которое эксперт не сможет отличить от поведения человека, обладающего определенными мыслительными способностями (скажем, способностью вьшблнять операции сложения или понимать китайский язык), то компьютер также обладает этими способностями. Следовательно, цель заключается лишь в том, чтобы создать программы, способные моделировать человеческое мышление таким образом, чтобы выдерживать тест Тьюринга. Более того, такая программа будет не просто моделью разума; она в буквальном смысле слова сама и будет разумом, в том же смысле, в котором человеческий разум — это разум.

Конечно, далеко не каждый специалист по искусственному интеллекту разделяет такую крайнюю точку зрения. Более осторожный подход заключается в том, чтобы рассматривать компьютерные модели как полезное средство для изучения разума, подобно тому как они применяются при изучении погоды, пищеварения, экономики или механизмов молекулярной биологии. Чтобы провести различие между этими двумя подходами, я назову первый «сильным ИИ», а второй — «слабым ИИ». Важно понять, насколько радикальным является подход сильного ИИ. Сильный ИИ утверждает, что мышление — это не что иное, как манипулирование формализованными символами, а именно это и делает компьютер: он оперирует формализованными символами. Подобный взгляд часто суммируется примерно следующим высказыванием: «Разум по отношению к мозгу — это то же, что и программа по отношению к аппаратуре компьютера».

СИЛЬНЫЙ ИИ отличается от других теорий разума по крайней мере в двух отношениях: его можно четко сформулировать, но также четко и просто его можно опровергнуть. Характер этого опровержения таков, что каждый человек может попробовать провести его самостоятельно. Вот как это делается. Возьмем, например, какой-нибудь язык, которого вы не понимаете. Для меня таким языком является китайский. Текст, написанный по-китайски, я воспринимаю как набор бессмысленных каракулей. Теперь предположим, что меня поместили в комнату, в которой расставлены корзинки, полные китайских иероглифов. Предположим также, что мне дали учебник на английском языке, в котором приводятся правила сочетания символов китайского языка, причем правила эти можно применять, зная лишь форму символов, понимать значение символов совсем необязательно. Например, правила могут гласить: «Возьмите такой-то иероглиф из корзинки номер один и поместите его рядом с таким-то иероглифом из корзинки номер два».



Я выдерживаю тест Тьюринга на понимание китайского языка

Представим себе, что находящиеся за дверью комнаты люди, понимающие китайский язык, передают в комнату наборы символов и что в ответ я манипулирую символами согласно правилам и передаю обратно другие наборы символов. В данном случае книга правил есть не что иное, как «компьютерная программа». Люди, написавшие ее, — «программисты», а я играю роль «компьютера». Корзинки, наполненные символами, — это «база данных»; наборы символов, передаваемых в комнату, это «вопросы», а наборы, выходящие из комнаты, это «ответы».

Предположим далее, что книга правил написана так, что мои «ответы» на «вопросы» не отличаются от ответов человека, свободно владеющего китайским языком. Например, люди, находящиеся снаружи, могут передать непонятные мне символы, означающие; «Какой цвет вам больше всего нравится?» В ответ, выполнив предписанные правилами манипуляции, я выдам символы, к сожалению, мне также непонятные и означающие, что мой любимый цвет синий, но мне также очень нравится зеленый. Таким образом, я выдержу тест Тьюринга на понимание китайского языка. Но все же на самом деле я не понимаю ни слова по-китайски. К тому же я никак не могу научиться этому языку в рассматриваемой системе, поскольку не существует никакого способа, с помощью которого я мог бы узнать смысл хотя бы одного символа. Подобно компьютеру, я манипулирую символами, но не могу придать им какого бы то ни было смысла.

Сущность этого мысленного эксперимента состоит в следующем: если я не могу понять китайского языка только потому, что выполняю компьютерную программу для понимания китайского, то и никакой другой цифровой компьютер не сможет его понять таким образом. Цифровые компьютеры просто манипулируют формальными символами согласно правилам, зафиксированным в программе.

То, что касается китайского языка, можно сказать и о других формах знания. Одного умения манипулировать символами еще недостаточно, чтобы гарантировать знание, восприятие, понимание, мышление и т. д. И поскольку компьютеры как таковые — это устройства, манипулирующие символами, наличия компьютерной программы недостаточно, чтобы можно было говорить о наличии знания.

Этот простой аргумент имеет решающее значение для опровержения концепции сильного ИИ. Первая предпосылка аргумента просто констатирует формальный характер компьютерной программы. Программы определяются в терминах манипулирования символами, а сами символы носят чисто формальный, или «синтаксический» характер. Между прочим, именно благодаря формальной природе программы, компьютер является таким мощным орудием. Одна и та же программа может выполняться на машинах самой различной природы, равно как одна и та же аппаратная система способна выполнять самые разнообразные компьютерные программы. Представим это соображение кратко в виде «аксиомы»:

Аксиома 1. *Компьютерные программы — это формальные (синтаксические) объекты.*

Это положение настолько важно, что его стоит рассмотреть несколько подробнее. Цифровой компьютер обрабатывает информацию, сначала кодируя ее в символических обозначениях, используемых в машине, а затем манипулируя символами в соответствии с набором строго определенных правил. Эти правила представляют собой программу. Например, в рамках тьюринговской концепции компьютера в роли символов выступали просто 0 и 1, а правила программы предписывали такие операции, как «Записать 0 на ленте, продвинуться на одну ячейку влево и стереть 1». Компьютеры обладают удивительным свойством: любая представимая на естественном языке информация может быть закодирована в такой системе обозначений и любая задача по обработке информации может быть решена путем применения правил, которые можно запрограммировать.

ВАЖНОЕ значение имеют еще два момента. Во-первых, символы и программы — это чисто абстрактные понятия: они не обладают физическими свойствами, с помощью которых их можно было бы определить и реализовать в какой бы то ни было физической среде. Нули и единицы, как символы, не имеют физических свойств. Я акцентирую на

этом внимание, поскольку иногда возникает соблазн отождествить компьютеры с той или иной конкретной технологией — скажем, с кремниевыми интегральными микросхемами — и считать, что речь идет о физических свойствах кремниевых кристаллов или что синтаксис означает какое-то физическое явление, обладающее, может быть, еще неизвестными каузальными свойствами аналогично реальным физическим явлениям, таким как электромагнитное излучение или атомы водорода, которые обладают физическими, каузальными свойствами. Второй момент заключается в том, что манипуляция символами осуществляется без всякой связи с каким бы то ни было смыслом. Символы в программе могут обозначать все, что угодно программисту или пользователю. В этом смысле программа обладает синтаксисом, но не обладает семантикой.

Следующая аксиома является простым напоминанием об очевидном факте, что мысли, восприятие, понимание и т. п. имеют смысловое содержание. Благодаря этому содержанию они могут служить отражением объектов и состояний реального мира. Если смысловое содержание связано с языком, то в дополнение к семантике, в нем будет присутствовать и синтаксис, однако лингвистическое понимание требует по крайней мере семантической основы. Если, например, я размышляю о последних президентских выборах, то мне в голову приходят определенные слова, но эти слова лишь потому относятся к выборам, что я придаю им специфическое смысловое значение в соответствии со своим знанием английского языка. В этом отношении они для меня принципиально отличаются от китайских иероглифов. Сформулируем это кратко в виде следующей аксиомы:

Аксиома 2. *Человеческий разум оперирует смысловым содержанием (семантикой).*

Теперь добавим еще один момент, который был продемонстрирован экспериментом с китайской комнатой. Располагать только символами как таковыми (т. е. синтаксисом) еще недостаточно для того, чтобы располагать семантикой. Простого манипулирования символами недостаточно, чтобы гарантировать знание их смыслового значения. Кратко представим это в виде аксиомы.

Аксиома 3. *Синтаксис сам по себе не составляет семантику и его недостаточно для существования семантики.*

На одном уровне этот принцип справедлив по определению. Конечно, кто-то может определить синтаксис и семантику по-иному. Главное, однако, в том, что существует различие между формальными элементами, не имеющими внутреннего смыслового значения, или содержания, и теми явлениями, у которых такое содержание есть. Из рассмотренных предпосылок следует:

Заключение 1. *Программы не являются сущностью разума и их наличия недостаточно для наличия разума.*



Компьютерные программы носят формальный характер (синтаксис). Человеческий разум имеет смысловое содержание (семантика).

.

А это по существу означает, что утверждение сильного ИИ ложно.

Очень важно отдавать себе отчет в том, что именно было доказано с помощью этого рассуждения и что нет.

Во-первых, я не пытался доказывать, что «компьютер не может мыслить». Поскольку все, что поддается моделированию вычислениями, может быть описано как компьютер, и поскольку наш мозг на некоторых уровнях поддается моделированию, то отсюда тривиально следует, что наш мозг — это компьютер, и он, разумеется, способен мыслить. Однако из того факта, что систему можно моделировать посредством манипулирования символами и что она способна мыслить, вовсе не следует, что способность к мышлению эквивалентна способности к манипулированию формальными символами.

Во-вторых, я не пытался доказывать, что только системы биологической природы, подобные нашему мозгу, способны мыслить. В настоящее время это единственные известные нам системы, обладающие такой способностью, однако мы можем встретить во Вселенной и другие способные к осознанным мыслям системы, а может быть, мы даже сумеем искусственно создать мыслящие системы. Я считают этот вопрос открытым для споров.

В-третьих, утверждение сильного ИИ заключается не в том, что компьютеры с правильными программами могут мыслить, что они могут обладать какими-то неведомыми доселе психологическими свойствами; скорее, оно состоит в том, что компьютеры просто должны мыслить, поскольку их работа — это и есть не что иное, как мышление.

В-четвертых, я попытался опровергнуть сильный ИИ, определенный именно таким образом. Я пытался доказать, что мышление не сводится к программам, потому что программа лишь манипулирует формальными символами — а, как нам известно, самого по себе манипулирования символами недостаточно, чтобы гарантировать наличие смысла. Это тот принцип, на котором основано рассуждение о китайской комнате.

Я подчеркиваю здесь эти моменты отчасти потому, что П.М. и П.С.Черчленды в своей статье (см. Пол М. Черчленд и Патриция Смит Черчленд «Может ли машина мыслить?»), как мне кажется, не совсем правильно поняли суть моих аргументов. По их мнению, сильный ИИ утверждает, что компьютеры в конечном итоге могут обрести способность к мышлению и что я отрицаю такую возможность, рассуждая лишь на уровне здравого смысла. Однако сильный ИИ утверждает другое, и мои доводы против не имеют ничего общего со здравым смыслом.

Далее я скажу еще кое-что об их возражениях. А пока я должен заметить, что в противоположность тому, что говорят Черчленды, рассуждение с китайской комнатой опровергает любые утверждения сильного ИИ относительно новых параллельных технологий, возникших под влиянием и моделирующих работу нейронных сетей. В отличие от компьютеров традиционной архитектуры фон Неймана, работающих в последовательном пошаговом режиме, эти системы располагают многочисленными вычислительными элементами, работающими параллельно и взаимодействующими друг с другом в соответствии с правилами, основанными на открытиях нейробиологии. Хотя пока достигнуты скромные результаты, модели «параллельной распределенной обработки данных» или «коммутационные машины» подняли некоторые полезные вопросы относительно того, насколько сложными должны быть параллельные системы, подобные нашему мозгу, чтобы при их функционировании порождалось разумное поведение.

Однако параллельный, «подобный мозгу» характер обработки информации не является существенным для чисто вычислительных аспектов процесса. Любая функция, которая может быть вычислена на параллельной машине, будет вычислена и на последовательной. И действительно, ввиду того что параллельные машины еще редки, параллельные программы обычно все еще выполняются на традиционных последовательных машинах. Следовательно, параллельная обработка также не избегает аргумента, основанного на примере с китайской комнатой.

Более того, параллельные системы подвержены своей специфической версии первоначального опровергающего рассуждения в случае с китайской комнатой. Вместо китайской комнаты представьте себе китайский спортивный зал, заполненный большим числом людей, понимающих только английский язык. Эти люди будут выполнять те же самые операции, которые выполняются узлами и синапсами в машине коннекционной архитектуры, описанной Черчленда-ми, но результат будет тем же, что и в примере с одним человеком, который манипулирует символами согласно правилам, записанным в руководстве. Никто в зале не понимает ни слова по-китайски, и не существует способа, следуя которому вся система в целом могла бы узнать о смысловом значении хотя бы одного китайского слова. Тем не менее при правильных инструкциях эта система способна правильно отвечать на вопросы, сформулированные по-китайски.

У параллельных сетей, как я уже говорил, есть интересные свойства, благодаря которым они могут лучше моделировать мозговые процессы по сравнению с машинами с традиционной последовательной архитектурой. Однако преимущества параллельной архитектуры, существенные для слабого ИИ, не имеют никакого отношения к противопоставлению между аргументом, построенным на примере с китайской комнатой, и утверждением сильного ИИ. Черчленды упускают из виду этот момент, когда говорят, что достаточно большой китайский спортивный зал мог бы обладать более высокими умственными способностями, которые определяются размерами и степенью сложности системы, равно как и мозг в целом более «разумен», чем его отдельные нейроны. Возможно и так, но это не имеет никакого отношения к вычислительному процессу. С точки зрения выполнения вычислений последовательные и параллельные архитектуры совершенно идентичны: любое вычисление, которое может быть произведено в машине с параллельным режимом работы, может быть выполнено машиной с последовательной архитектурой. Если человек, находящийся в китайской комнате и производящий вычисления эквивалентен и той и другой системам, тогда, если он не понимает китайского языка исключительно потому, что ничего кроме вычислений не делает, то и эти системы также не понимают китайского языка. Черчленды правы, когда говорят, что первоначальный довод, основанный на примере с китайской комнатой, был сформулирован исходя из традиционного представления об ИИ, но они заблуждаются, считая что параллельная архитектура делает этот довод неуязвимым. Это справедливо в отношении любой вычислительной системы. Производя только формальные операции с символами (т. е. вычисления) вы не сможете обогатить свой разум семантикой, независимо от того выполняются эти вычислительные операции последовательно или параллельно; вот почему аргумент китайской комнаты опровергает сильный ИИ в любой его форме.

МНОГИЕ люди, на которых этот аргумент производит определенное впечатление, тем не менее затрудняются провести четкое различие между людьми и компьютерами.

Если люди, по крайней мере в тривиальном смысле, являются компьютерами и если люди обладают семантикой, то почему они не могут наделить семантикой и другие компьютеры? Почему мы не можем запрограммировать компьютеры Vax или Cray таким образом, чтобы у них тоже появились мысли и чувства? Или почему какая-нибудь новая компьютерная технология не сможет преодолеть пропасть, разделяющую форму и содержание, или синтаксис и семантику? В чем на самом деле состоит то различие между биологическим мозгом и компьютерной системой, благодаря которому аргумент с китайской комнатой действует применительно к компьютерам, но не действует применительно к мозгу?

Наиболее очевидное различие заключается в том, что процессы, которые определяют нечто как компьютер (а именно вычислительные процессы), на самом деле совершенно не зависят от какого бы то ни было конкретного типа аппаратной реализации. В принципе можно сделать компьютер из старых жестяных банок из-под пива, соединив их проволокой и обеспечив энергией от ветряных мельниц.

Однако когда мы имеем дело с мозгом, то хотя современная наука в значительной степени еще пребывает в неведении относительно протекающих в мозгу процессов, мы поражаемся чрезвычайной специфичности анатомии и физиологии. Там, где мы достигли некоторого понимания того, как мозговые процессы порождают те или иные психические явления, — например, боль, жажду, зрение, обоняние — нам ясно, что в этих процессах участвуют вполне определенные нейробиологические механизмы. Чувство жажды, по крайней мере в некоторых случаях, обусловлено срабатыванием нейронов определенных типов в гипоталамусе, которое в свою очередь вызвано действием специфического пептида, ангиотензина II. Причинные связи прослеживаются здесь «снизу вверх» в том смысле, что нейронные процессы низшего уровня обусловливают психические явления на более высоких уровнях. В самом деле, каждое «ментальное» явление, от чувства жажды до мыслей о математических теоремах и воспоминаний о детстве, вызывается срабатыванием определенных нейронов в определенных нейронных структурах.

Однако почему эта специфичность имеет такое важное значение? В конце концов всевозможные срабатывания нейронов можно смоделировать на компьютерах, физические и химические свойства которых совершенно отличны от свойств мозга. Ответ состоит в том, что мозг не просто демонстрирует формальные процедуры или программы (он делает и это тоже), но и вызывает ментальные события благодаря специфическим нейробиологическим процессам. Мозг по сути своей является биологическим органом и именно его особые биохимические свойства позволяют достичь эффекта сознания и других видов ментальных явлений. Компьютерные модели мозговых процессов обес-

печивают отражение лишь формальных аспектов этих процессов. Однако моделирование не следует смешивать с воспроизведением. Вычислительные модели ментальных процессов не ближе к реальности, чем вычислительные модели любого другого природного явления.

Можно представить себе компьютерную модель, отражающую воздействие пептидов на гипоталамус, которая будет точна вплоть до каждого отдельного синапса. Но с таким же успехом мы можем представить себе компьютерное моделирование процесса окисления углеводородов в автомобильном двигателе или пищеварительного процесса в желудке. И модель процессов, протекающих в мозге, ничуть не реальнее моделей, описывающих процессы сгорания топлива или пищеварительные процессы. Если не говорить о чудесах, то вы не сможете привести свой автомобиль в движение, моделируя на компьютере окисление бензина, и вы не сможете переварить обед, выполняя программу, которая моделирует пищеварение. Представляется очевидным и тот факт, что и моделирование мышления также не произведет нейробиологического эффекта мышления.

Следовательно, все ментальные явления вызываются нейробиологическими процессами мозга. Представим сокращенно этот тезис следующим образом:

Аксиома 4. *Мозг порождает разум.*

В соответствии с рассуждениями, приведенными выше, я немедленно прихожу к тривиальному следствию.

Заключение 2. *Любая другая система, способная порождать разум, должна обладать каузальными свойствами (по крайней мере), эквивалентными соответствующим свойствам мозга.*

Это равносильно, например, следующему утверждению: если электрический двигатель способен обеспечивать автомашине такую же высокую скорость, как двигатель внутреннего сгорания, то он должен обладать (по крайней мере) эквивалентной мощностью. В этом заключении ничего не говорится о механизмах. На самом деле, мышление — это биологическое явление: психические состояния и процессы обусловлены процессами мозга. Из этого еще не следует, что только биологическая система может мыслить, но это в то же время означает, что любая система другой природы, основанная на кремниевых кристаллах, жестяных банках и т. п., должна будет обладать каузальными возможностями, эквивалентными соответствующим возможностям мозга. Таким образом, я прихожу к следующему выводу:

Заключение 3. *Любой артефакт, порождающий ментальные явления, любой искусственный мозг должен иметь способность вопроизводить специфические каузальные свойства мозга, и наличия этих свойств невозможно добиться только за счет выполнения формальной программы.*

Более того, я прихожу к важному выводу, касающемуся человеческого мозга:

Заключение 4. *Тот способ, посредством которого человеческий мозг на самом деле порождает ментальные явления, не может сводиться лишь к выполнению компьютерной программы.*



Кому могло прийти в голову, что компьютерное моделирование процесса мышления и сам мыслительный (ментальный процесс) — это одно и то же?

.

ВПЕРВЫЕ сравнение с китайской комнатой было приведено мною на страницах журнала "Behavioral and Brain Sciences" (Науки о поведении и мозге) в 1980 г. Тогда моя статья сопровождалась, в соответствии с принятой в этом журнале практикой, коммен-

тариями оппонентов, в данном случае свои соображения высказали 26 оппонентов. Откровенно говоря, мне кажется, что смысл этого сравнения довольно очевиден, но, к моему удивлению, статья и в дальнейшем вызвала целый поток возражений, и что еще более удивительно, этот поток продолжается и по сей день. По-видимому, аргумент китайской комнаты затронул какое-то очень больное место.

Основной тезис сильного ИИ заключается в том, что любая система (независимо, сделана ли она из пивных банок, кремниевых кристаллов или просто из бумаги) не только способна обладать мыслями и чувствами, но просто *должна* ими обладать, если только она реализует правильно составленную программу, с правильными входами и выходами. Очевидно, это абсолютно антибиологическая точка зрения, и естественно было бы ожидать, что специалисты по искусственному интеллекту охотно откажутся от нее. Многие из них, особенно представители молодого поколения, согласны со мной, но меня поражает, как много сторонников имеет эта точка зрения и как настойчиво они защищают ее. Приведу некоторые наиболее распространенные из высказываемых ими доводов:

а) В китайской комнате вы на самом деле понимаете китайский, хотя и не отдаете себе в этом отчета. В конце концов вы можете понимать что-то и не отдавая себе в этом отчета.

б) Вы не понимаете китайского, но в вас существует подсистема (подсознание), которая понимает. Существуют ведь подсознательные психические состояния, и нет причины считать, что ваше понимание китайского не могло бы быть полностью неосознанным.

в) Вы не понимаете китайского, но комната как целое — понимает. Вы подобны отдельному нейрону в мозгу, и нейрон как таковой не может ничего понимать, он лишь вносит свой вклад в то понимание, которое демонстрирует система в целом; вы сами не понимаете, но вся система понимает.

г) Никакой семантики и не существует: есть только синтаксис. Полагать, что в мозгу есть какое-то загадочное «психическое содержание», «мыслительные процессы» или «семантика», это своего рода донаучная иллюзия. Все, что на самом деле существует в мозгу, — это некоторое синтаксическое манипулирование символами, которое осуществляется и в компьютерах. И ничего больше.

д) В действительности вы не выполняете компьютерную программу — это вам только кажется. Если существует некий сознательный агент, следующий строкам программы, то процесс уже вовсе не является простой реализацией программы.

е) Компьютеры обладали бы семантикой, а не только синтаксисом, если бы их входы и выходы были поставлены в причинные, каузальные зависимости — по отношению к

остальному миру. Допустим, что мы снабдили робота компьютером, подключили телевизионные камеры к его голове, установили трансдьюсеры, подводящие телевизионную информацию к компьютеру, и позволили последнему управлять руками и ногами робота. В таком случае система как целое будет обладать семантикой.

ж) Если программа моделирует поведение человека, говорящего по-китайски, то она понимает китайский язык. Предположим, что нам удалось смоделировать работу мозга китайца на уровне нейронов. Но тогда, конечно, подобная система будет понимать китайский так же хорошо, как и мозг любого китайца.

И так далее.

У всех этих доводов есть одно общее свойство: все они неадекватны рассматриваемой проблеме, потому что не улавливают самой сути рассуждения о китайской комнате. Эта суть заключается в различии между формальным манипулированием символами, осуществляемым компьютером, и смысловым содержанием, биологически порождаемым мозгом, — различии, которое я для краткости выражения (и надеюсь, что это никого не введет в заблуждение) свел к различию между синтаксисом и семантикой. Я не буду повторять своих ответов на все эти возражения, однако проясню, возможно, ситуацию, если скажу, в чем заключаются слабости наиболее распространенного довода моих оппонентов, а именно довода (в), который я назову ответом системы. (Очень часто встречается также и довод (ж), основанный на идее моделирования мозга, но он уже был рассмотрен выше.)

В ОТВЕТЕ системы утверждается, что *вы,* конечно, не понимаете китайского, но вся система в целом — вы сами, комната, свод правил, корзинки, наполненные символами, — понимает. Когда я впервые услышал это объяснение, я спросил высказавшего это объяснение человека: «Вы что же, считаете, что комната может понимать китайский язык?» Он ответил, да. Это, конечно, смелое утверждение, однако, помимо того что оно совершенно неправдоподобно, оно не состоятельно еще и с чисто логической точки зрения. Суть моего исходного аргумента была в том, что простое тасование символов еще не обеспечивает доступа к пониманию смысла этих символов. Но это в той же мере касается комнаты в целом, как и находящегося в ней человека. В правоте моих слов можно убедиться, несколько расширив наш мысленный эксперимент. Представим себе, что я заучил наизусть содержимое корзинок и книги правил и что я провожу все вычисления в уме. Допустим даже, что я работаю не в комнате, а у всех на виду. В системе не осталось ничего такого, чего бы не было во мне самом, но поскольку я не понимаю китайского языка, не понимает его и система.

В своей статье мои оппоненты Черчленды используют одну из разновидностей ответа системы, придумав любопытную аналогию. Предположим, кто-то стал утверждать, что

свет не может иметь электромагнитную природу, поскольку, когда человек перемещает магнит в темной комнате, мы не наблюдаем видимого светового излучения. Приведя этот пример, Черчленды спрашивают, а не является ли аргумент с китайской комнатой чем-то в том же роде? Не равносильно ли будет сказать, что когда вы манипулируете китайскими иероглифами в семантически темной комнате, в ней не возникает никакого просвета в понимании китайского языка? Но не может ли потом в ходе будущих исследований выясниться — так же, как было доказано, что свет все-таки целиком состоит из электромагнитного излучения, — что семантика целиком и полностью состоит из синтаксиса? Не является ли этот вопрос предметом дальнейшего научного изучения?

Аргументы, построенные на аналогиях, всегда очень уязвимы, поскольку, прежде чем аргумент станет состоятельным, необходимо еще убедиться, что две рассматриваемые ситуации действительно аналогичны. В данном случае, я думаю, что это не так. Объяснение света на основе электромагнитного излучения — это причинное рассуждение от начала и до конца. Это причинное объяснение физики электромагнитных волн. Однако аналогия с формальными символами не состоятельна, поскольку формальные символы не имеют физических причинных свойств. Единственное, что во власти символов как таковых, — это вызвать следующий шаг в программе, которую выполняет работающая машина. И здесь не возникает никакой речи о дальнейших исследованиях, которым еще предстоит раскрыть доселе неизвестные физические причинные свойства нулей и единиц. Последние обладают лишь одним видом свойств — абстрактными вычислительными свойствами, которые уже хорошо изучены.

Черчленды говорят, что у них «напрашивается вопрос», когда я утверждаю, что интерпретированные формальные символы не идентичны смысловому содержанию. Да, я, конечно, не тратил много времени на доказательство, что это так, поскольку я считаю это логической истиной. Как и в случае с любой другой логической истиной, каждый может быстро убедиться, что она справедлива, поскольку, предположив обратное, сразу приходишь к противоречию. Попробуем провести такое доказательство. Предположим, что в китайской комнате имеет место какое-то скрытое понимание китайского языка. Что же может превратить процесс манипулирования синтаксическими элементами в специфично китайское смысловое содержание? Подумав, я в конце концов пришел к выводу, что программисты должны были говорить по-китайски, коль скоро они сумели запрограммировать систему для обработки информации, представленной на китайском языке.

Хорошо. Но теперь представим себе, что надоело, сидя в китайской комнате, тасовать китайские (для меня бессмысленные) символы. Предположим, мне пришло в голову интерпретировать эти символы как обозначения ходов в шахматной игре. Какой семан-

тикой теперь обладает система? Обладает ли она китайской семантикой или шахматной, или она обладает одновременно и той и другой? Предположим, что есть еще некая третья личность, наблюдающая за мной в окошко, и она решает, что мое манипулирование символами можно интерпретировать как предсказание курса акций на бирже. И так далее. Не существует предела количеству семантических интерпретаций, которое можно приписать символам, поскольку, я повторяю, символы носят чисто формальный характер. Они не содержат в себе внутренней семантики.

Можно ли каким-то образом спасти аналогию Черчлендов? Выше я сказал, что формальные символы не имеют каузальных свойств. Но, конечно, программа всегда выполняется той или иной конкретной аппаратурой, и эта аппаратура обладает своими специфическими физическими, каузальными свойствами. Любой реальный компьютер порождает различные физические явления. Мой компьютер, к примеру, выделяет тепло, производит монотонный шум и т. д. Существует ли здесь какое-либо строгое логическое доказательство, что компьютер не может производить аналогичным образом эффект сознания? Нет. В научном смысле об этом и речи быть не может, однако это совсем не то, что призвано опровергать рассуждение о китайской комнате, и не то, на чем будут настаивать сторонники сильного ИИ, поскольку любой производимый таким образом эффект будет достигать за счет физических свойств реализующей программу среды. Основное утверждение сильного ИИ заключается в том, что физические свойства реализующей среды не имеют никакого значения. Имеют значение лишь программы, а программы — это чисто формальные объекты.

Таким образом аналогия Черчлендов между синтаксисом и электромагнитным излучением наталкивается на дилемму: либо синтаксис следует понимать чисто формально, через его абстрактные математические свойства, либо нет. Если выбрать первую альтернативу, то аналогия становится несостоятельной, поскольку синтаксис, понимаемый таким образом, не имеет физических свойств. Если же, с другой стороны, рассматривать синтаксис в плоскости физических свойств реализующей среды, тогда аналогия действительно состоятельна, но она не имеет отношения к сильному ИИ.

ПОСКОЛЬКУ сделанные мною утверждения довольно очевидны — синтаксис это не то же самое, что семантика; мозговые процессы порождают психические явления — возникает вопрос, а как вообще возникла эта путаница? Кому могло прийти в голову, что компьютерное моделирование ментального процесса полностью ему идентично? В конце концов весь смысл моделей заключается в том, что они улавливают лишь какую-то часть моделируемого явления и не затрагивают остального. Ведь никто не думает, что мы-захотим поплавать в бассейне, наполненном шариками для пинг-понга, моделирующими молекулы воды. Можно ли тогда считать, что компьютерная модель мыслительных процессов на самом деле способна мыслить?

Отчасти эти недоразумения объясняются тем, что люди унаследовали некоторые положения -бихевиористских психологических теорий прошлого поколения. Под тестом Тьюринга скрывается соблазн считать, что если нечто ведет себя так, как будто оно обладает ментальными процессами, то оно и на самом деле должно ими обладать. Частью ошибочной бихевиористской концепции было также и то, что психология, для того чтобы оставаться научной дисциплиной, должна ограничиваться изучением внешне наблюдаемого поведения. Парадоксально, но этот остаточный бихевиоризм связан с остаточным дуализмом. Никто не думает, что компьютерная модель пищеварения способна что-то переварить на самом деле, но там, где речь идет о мышлении, люди охотно верят в такие чудеса, потому что забывают о том, что разум — это такое же биологическое явление, как и пищеварение. По их мнению, разум — это нечто формальное и абстрактное, а вовсе не часть полужидкой субстанции, из которой состоит наш головной мозг. Полемическая литература по искусственному интеллекту обычно содержит нападки на то, что авторы называют дуализмом, но при этом они не замечают, как сами демонстрируют ярко выраженный дуализм, поскольку если не принять точку зрения, что разум совершенно не зависит от мозга или какой-либо другой физически специфической системы, то следует считать невозможным создание разума только за счет написания программ.

Исторически в странах Запада научные концепции, в которых люди рассматривались как часть обычного физического или биологического мира, часто встречали противодействие со стороны реакции. Идеям Коперника и Галилея противились, потому что они отрицали, что Земля является центром Вселенной. Против Дарвина выступали потому, что он утверждал, что люди произошли от низших животных. Сильный ИИ правильнее всего было бы рассматривать как одно из последних проявлений этой антинаучной традиции, так как он отрицает, что человеческий разум содержит что-то существенно физическое или биологическое. Согласно утверждениям сильного ИИ, разум не зависит от мозга. Он представляет собой компьютерную программу и по существу не связан ни с какой специфической аппаратурой.

Многие люди, сомневающиеся относительно физической значимости искусственного интеллекта, полагают, что компьютеры, может быть, и смогут понимать китайский язык или думать о числах, но принципиально не способны на проявления чисто человеческих свойств, а именно (и далее следует их излюбленная человеческая специфика): любовь, чувство юмора, тревога за судьбу постиндустриального общества в эпоху современного капитализма и т. д. Но специалисты по ИИ справедливо настаивают, что эти возражения не корректны, что здесь как бы отодвигаются футбольные ворота. Если искусственное моделирование интеллекта окажется успешным, то психологические вопросы уже не имеют сколь-нибудь важного значения, В этом споре обе стороны не за-

мечают различия между моделированием и воспроизведением. Пока речь идет о моделировании, то не стоит никакого труда запрограммировать мой компьютер, чтобы он напечатал: «Я люблю тебя, Сюзи»; «Ха-ха!» или «Я испытываю тревоги постиндустриального общества». Важно отдавать себе отчет в том, что моделирование — это не то же самое, что воспроизведение; и этот факт имеет такое же отношение к размышлениям об арифметике, как и к чувству тревоги. Дело не в том, что компьютер доходит только до центра поля и не доходит до ворот. Компьютер даже не трогается с места. Он просто не играет в эту игру.

Искусственный интеллект:
Может ли машина мыслить?

Классический искусственный интеллект едва ли будет воплощен в мыслящих машинах; предел человеческой изобретательности в этой области, по-видимому, ограничится созданием систем, имитирующих работу мозга

ПОЛ М. ЧЕРЧЛЕНД, ПАТРИЦИЯ СМИТ ЧЕРЧЛЕНД

В НАУКЕ об искусственном интеллекте (ИИ) происходит революция. Чтобы объяснить ее причины и смысл и представить рассуждения Джона Р. Сирла в перспективе, мы прежде должны обратиться к истории.

В начале 50-х годов традиционный, несколько расплывчатый вопрос о том, может ли машина мыслить, уступил более доступному вопросу: может ли мыслить машина, манипулирующая физическими символами в соответствии с правилами, учитывающими их структуру. Этот вопрос сформулирован точнее, потому что за предшествовавшие полвека формальная логика и теория вычислений существенно продвинулись вперед. Теоретики стали высоко оценивать возможности абстрактных систем символов, которые претерпевают преобразования в соответствии с определенными правилами. Казалось, что если эти системы удалось бы автоматизировать, то их абстрактная вычислительная мощь проявилась бы в реальной физической системе. Подобные взгляды способствовали рождению вполне определенной программы исследований на достаточно глубокой теоретической основе.

Может ли машина мыслить? Было много причин для того, чтобы ответить да. Исторически одна из первых и наиболее глубоких причин заключалась в двух важных результатах теории вычислений. Первый результат был тезисом Черча, согласно которому каждая эффективно вычислимая функция является рекурсивно вычислимой. Термин «эффективно вычислимая» означает, что существует некая «механическая» процедура,

с помощью которой можно за конечное время вычислить результат при заданных входных данных. «Рекурсивно вычислимая» означает, что существует конечное множество операций, которые можно применить к заданному входу, а затем последовательно и многократно применять к вновь получаемым результатам, чтобы вычислить функцию за конечное время. Понятие механической процедуры не формальное, а скорее интуитивное, и потому тезис Черча не имеет формального доказательства. Однако он проникает в самую суть того, чем является вычисление, и множество различных свидетельств сходится в его подтверждение.

Второй важный результат был получен Аланом М. Тьюрингом, продемонстрировавшим, что любая рекурсивно вычислимая функция может быть вычислена за конечное время с помощью максимально упрощенной машины, манипулирующей символами, которую позднее стали называть универсальной машиной Тьюринга. Эта машина управляется рекурсивно применимыми правилами, чувствительными к идентичности, порядку и расположению элементарных символов, которые играют роль входных данных.

ИЗ ЭТИХ двух результатов вытекает очень важное следствие, а именно что стандартный цифровой компьютер, снабженный правильной программой, достаточно большой памятью и располагающий достаточным временем, может вычислить *любую* управляемую правилами функцию с входом и выходом. Другими словами, он может продемонстрировать любую систематическую совокупность ответов на произвольные воздействия со стороны внешней среды.

Конкретизируем это следующим образом: рассмотренные выше результаты означают, что соответственно запрограммированная машина, манипулирующая символами (в дальнейшем будем называть ее МС-машиной), должна удовлетворять тесту Тьюринга на наличие сознательного разума. Тест Тьюринга — это чисто бихевиористский тест, тем не менее его требования очень сильны. (Насколько состоятелен этот тест, мы рассмотрим ниже, там где встретимся со вторым, принципиально отличным «тестом» на наличие сознательного разума.) Согласно первоначальной версии теста Тьюринга, входом для МС-машины должны быть вопросы и фразы на естественном разговорном языке, которые мы набираем на клавиатуре устройства ввода, а выходом являются ответы МС-машины, напечатанные устройством вывода. Считается, что машина выдержала этот тест на присутствие сознательного разума, если ее ответы невозможно отличить от ответов, напечатанных реальным, разумным человеком. Конечно, в настоящее время никому не известна та функция, с помощью которой можно было бы получить выход, не отличающийся от поведения разумного человека. Но результаты Черча и Тьюринга гарантируют нам, что какова бы ни была эта (предположительно эффективная) функция, МС-машина соответствующей конструкции сможет ее вычислить.

Это очень важный вывод, особенно если учесть, что тьюринговское описание взаимодействия с машиной при помощи печатающей машинки представляет собой несущественное ограничение. То же заключение остается в силе, даже если МС-машина взаимодействует с миром более сложными способами: с помощью аппарата непосредственного зрения, естественной речи и т. д. В конце концов более сложная рекурсивная функция все же остается вычислимой по Тьюрингу. Остается лишь одна проблема: найти ту несомненно сложную функцию, которая управляет ответными реакциями человека на воздействия со стороны внешней среды, а затем написать программу (множество рекурсивно применимых правил), с помощью которой МС-машина вычислит эту функцию. Вот эти цели и легли в основу научной программы классического искусственного интеллекта.

Первые результаты были обнадеживающими. МС-машины с остроумно составленными программами продемонстрировали целый ряд действий, которые как будто относятся к проявлениям разума. Они реагировали на сложные команды, решали трудные арифметические, алгебраические и тактические задачи, играли в шашки и шахматы, доказывали теоремы и поддерживали простой диалог. Результаты продолжали улучшатьсгя с появлением более емких запоминающих устройств, более быстродействующих машин, а также с разработгюй более мощных и изощренных программ. Классический, или «построенный на программировании», ИИ представлял собой очень живое и успешное научное направление почти со всех точек зрения. Периодически высказывавшееся отрицание того, что МС-машины в конечном итоге будут способны мыслить, казалось проявлением необъективности и неинформированности. Свидетельства в пользу положительного ответа на вопрос, вынесенный в заголовок статьи, казались более чем убедительными.

Конечно, оставались кое-какие неясности. Прежде всего МС-машины не очень-то напоминали человеческий мозг. Однако и здесь у классического ИИ был наготове убедительный ответ. Во-первых, физический материал, из которого сделана МС-машина, по существу не имеет никакого отношения к вычисляемой ею функции. Последняя зафиксирована в программе. Во-вторых, техничекие подробности функциональной архитектуры машины также не имеют значения, поскольку совершенно различные архитектуры, рассчитанные на работу с совершенно различными программами, могут тем не менее выполнять одну и ту же функцию по входу-выходу.

Поэтому целью ИИ было найти функцию, по входу и выходу характерную для разума, а также создать наиболее эффективную из многих возможных программ для того, чтобы вычислить эту функцию. При этом говорили, что тот специфичный способ, с помощью

которого функция вычисляется человеческим мозгом, не имеет значения. Этим и завершается описание сущности классического ИИ и оснований для положительного ответа на вопрос, поставленный в заголовке статьи.

МОЖЕТ ЛИ машина мыслить? Были также кое-какие доводы и в пользу отрицательного ответа. На протяжении 60-х годов заслуживающие внимания отрицательные аргументы встречались относительно редко. Иногда высказывалось возражение, заключающееся в том, что мышление — это не физический процесс и протекает он в нематериальной душе. Однако подобное дуалистическое воззрение не выглядело достаточно убедительным ни с эволюционной, ни с логической точки зрения. Оно не оказало сдерживающего влияния на исследования в области ИИ.

Гораздо большее внимание специалистов по ИИ привлекли соображения иного характера. В 1972 г. Хьюберт Л. Дрейфус опубликовал книгу, в которой резко критиковались парадные демонстрации проявлений разума у систем ИИ. Он указывал на то, что эти системы не адекватно моделировали подлинное мышление, и вскрыл закономерность, присущую всем этим неудачным попыткам. По его мнению, в моделях отсутствовал тот огромный запас неформализованных общих знаний о мире, которым располагает любой человек, а также способность, присущая здравому рассудку, опираться на те или иные составляющие этих знаний,в зависимости от требований изменяющейся обстановки. Дрейфус не отрицал принципиальной возможности создания искусственной физичесгсой системы, способной мыслить, но он весьма критически отнесся к идее о том, что это может быть достигнуто только за счет манипулирования символами с помощью рекурсивно применяемых правил.

В кругах специалистов по искусственному интеллекту, а также философов рассуждения Дрейфуса были восприняты главным образом как недальновидные и необъективные, базирующиеся на неизбежных упрощениях, присущих этой еще очень молодой области исследований. Возможно, указанные недостатки действительно имели место, но они, конечно, были временными. Настанет время, когда более мощные машины и более качественные программы позволят избавиться от этих недостатков. Казалось, что время работает на искусственный интеллект. Таким образом, и эти возражения не оказали сколько-нибудь заметного влияния на дальнейшие исследования в области ИИ.

Однако оказалось, что время работало и на Дрейфуса: в конце 70-х - начале 80-х годов увеличение быстродействия и объема памяти компьютеров повышало их «умственные способности» ненамного. Выяснилось, например, что распознавание образов в системах машинного зрения требует неожиданно большого объема вычислений. Для получения практически достоверных результатов нужно было затрачивать все больше и больше машинного времени, намного превосходя время, требуемое для выполнения тех же задач биологической системе зрения. Столь медленный процесс моделирования

настораживал: ведь в компьютере сигналы распространяются приблизительно в миллион раз быстрее, чем в мозге, а тактовая частота центрального процессорного устройства компьютера примерно во столько же раз выше частоты любых колебаний, обнаруженных в мозге. И все же на реалистических задачах черепаха легко обгоняет зайца.

Кроме того, для решения реалистических задач необходимо, чтобы компьютерная программа обладала доступом к чрезвычайно обширной базе данных. Построение такой базы данных уже само по себе представляет довольно сложную проблему, но она усугубляется еще одним обстоятельством: каким образом обеспечить доступ к конкретным, зависящим от контекста фрагментам этой базы данных в реальном масштабе времени. По мере того как базы данных становились все более емкими, проблема доступа осложнялась. Исчерпывающий поиск занимал слишком много времени, а эвристические методы не всегда приводили к успеху. Опасения, подобные тем, что высказывал Дрейфус, начали разделять даже некоторые специалисты, работающие в области искусственного интеллекта.

Приблизительно в это время (1980 г.) Джон Сирл высказал принципиально новую критическую концепцию, ставившую под сомнение само фундаментальное предположение классической программы исследований по ИИ, а именно - идею о том, что правильное манипулирование структурированными символами путем рекурсивного применения правил, учитывающих их структуру, может составлять сущность сознательного разума.

Основной аргумент Сирла базировался на мысленном эксперименте, в котором он демонстрирует два очень важных обстоятельства. Во-первых, он описывает МС-машину, которая (как мы должны понимать) реализует функцию, по входу и выходу способную выдержать тест Тьюринга в виде беседы, протекающей исключительно на китайском языке. Во-вторых, внутренняя структура машины такова, что независимо от того, какое поведение она демонстрирует, у наблюдателя не возникает сомнений в том, что ни машина в целом, ни любая ее часть не понимают китайского языка. Все, что она в себе содержит, - это говорящий только по-английски человек, выполняющий записанные в инструкции правила, с помощью которых следует манипулировать символами, входящими и выходящими через окошко для почтовой корреспонденции в двери. Короче говоря, система положительно удовлетворяет тесту Тьюринга, несмотря на то что не обладает подлинным пониманием китайского языка и реального семантического содержания сообщений (см. статью Дж. Сирла «Разум мозга - компьютерная программа?»).

Отсюда делается общий вывод, что любая система, просто манипулирующая физическими символами согласно чувствительным к структуре правилам, будет в лучшем случае лишь жалкой пародией настоящего сознательного разума, поскольку невозможно

породить «реальную семантику», просто крутя ручку «пустого синтаксиса». Здесь следует заметить, что Сирл выдвигает не бихевиористский (не поведенческий) тест на наличие сознания: элементы сознательного разума должны обладать реальным семантическим содержанием.

Возникает соблазн упрекнуть Сирла в том, что его мысленный эксперимент не адекватен, поскольку предлагаемая им система, действующая по типу «кубика-рубика», будет работать до абсурда медленно. Однако Сирл настаивает, что быстродействие в данном случае не играет никакой роли. Думающий медленно все же думает верно. Все необходимое для воспроизведения мышления, согласно концепции классического ИИ, по его мнению, присутствует в «китайской комнате».

Статья Сирла вызвала оживленные отклики специалистов по ИИ, психологов и философов. Однако в общем и целом она была встречена еще более враждебно, чем книга Дрейфуса. В своей статье, которая одновременно публикуется в этом номере журнала, Сирл приводит ряд критических доводов, высказываемых против его концепции. По нашему мнению, многие из них правомерны, в особенности те, авторы которых жадно «кидаются на приманку», утверждая, что, хотя система, состоящая из комнаты и ее содержимого, работает ужасно медленно, она все же понимает китайский язык.

Нам нравятся эти ответы, но не потому, что мы считаем, будто китайская комната понимает китайский язык. Мы согласны с Сирлом, что она его не понимает. Привлекательность этих аргументов в том, что они отражают отказ воспринять важнейшую третью аксиому в рассуждении Сирла: « *Синтаксис сам по себе не составляет семантику и его недостаточно для существования семантики».* Возможно, эта аксиома и справедлива, но Сирл не может с полным основанием утверждать, что ему это точно известно. Более того, предположить, что она справедлива, - значит напрашиваться на вопрос о том, состоятельна ли программа исследований классического ИИ, поскольку эта программа базируется на очень интересном предположении, что если нам только удастся привести в движение соответствующим образом структурированный процесс, своеобразный внутренний танец синтаксических элементов, правильно связанный со входами и выходами, то мы можем получить те же состояния и проявления разума, которые присущи человеку.

То, что третья аксиома Сирла действительно напрашивается на этот вопрос, становится очевидно, когда мы непосредственно сопоставляем ее с его же первым выводом: *«Программы появляются сущностью разума и их наличия не достаточно для наличия разума».* Не трудно видеть, что его третья аксиома уже несет в себе 90% почти идентичного ей заключения. Вот почему мысленный эксперимент Сирла специально придуман для того, чтобы подкрепить третью аксиому. В этом вся суть китайской комнаты.

Хотя пример с китайской комнатой делает аксиому 3 привлекательной для непосвященного, мы не думаем, что он доказывает справедливость этой аксиомы, и чтобы продемонстрировать несостоятельность этого примера, мы предлагаем в качестве иллюстрации свой параллельный пример. Часто один удачный пример, опровергающий оспариваемое утверждение, значительно лучше проясняет ситуацию, чем целая книга, полная логического жонглирования.

В истории науки было много примеров скепсиса, подобного тому, который мы видим в рассуждениях Сирла. В XVIII в. ирландский епископ Джордж Беркли считал немыслимым, чтобы волны сжатия в воздухе сами по себе могли быть сущностью звуковых явлений или фактором, достаточным для их существования. Английский поэт и художник Уильям Блейк и немецкий поэт-естествоиспытатель Иоган Гете считали немыслимым, чтобы маленькие частички материи сами по себе могли быть сущностью или фактором, достаточным для объективного существования света. Даже в нынешнем столетии находились люди, которые не могли себе вообразить, чтобы неодушевленная материя сама по себе, независимо от того, насколько сложна ее организация, могла быть органической сущностью или достаточным условием жизни. Совершенно очевидно то, что люди могут или не могут себе представить, зачастую никак не связано с тем, что на самом деле существует или не существует в действительности. Это справедливо, даже когда речь идет о людях с очень высоким уровнем интеллекта.

Чтобы увидеть, каким образом эти исторические уроки можно применить к рассуждениям Сирла, применим искусственно придуманную параллель к его логике и подкрепим эту параллель мысленным экспериментом.

Аксиома 1. *Электричество и магнетизм - это физические силы.*

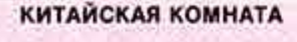
Аксиома 2. *Существенное свойство света - это свечение.*

Аксиома 3. *Силы сами по себе появляются сущностью эффекта свечения и не достаточны для его наличия.*

Заключение 1. *Электричество и магнетизм не являются сущностью света и не достаточны для его наличия.*

Предположим, что это рассуждение было опубликовано вскоре после того, как Джеймс К. Максвелл в 1864 г. высказал предположение, что свет и электромагнитные волны идентичны, но до того как в мире были полностью осознаны систематические параллели между свойствами света и свойствами электромагнитных волн. Приведенное выше логическое рассуждение могло показаться убедительным возражением против смелой гипотезы Максвелла, в особенности если бы оно сопровождалось следующим комментарием в поддержку аксиомы 3.

«Рассмотрим темную комнату, в которой находится человек, держащий в руках постоянный магнит или заряженный предмет. Если человек начнет перемещать магнит вверх-вниз, то, согласно теории Максвелла об искусственном освещении (ИО), от магнита будет исходить распространяющаяся сфера электромагнитных волн и в комнате станет светлее. Но, как хорошо известно всем, кто пробовал играть с магнитами или заряженными шарами, их силы (а если на то пошло, то и любые другие силы), даже когда эти объекты приходят в движение, не создают никакого свечения. Поэтому представляется немыслимым, чтобы мы могли добиться реального эффекта свечения просто за счет манипулирования силами!»



КИТАЙСКАЯ КОМНАТА

Аксиома 1. Компьтерные программы — это формальные (синтаксические) объекты.

Аксиома 2. Человеческий разум обладает смысловым содержанием (семантикой).

Аксиома 3. Синтаксис сам по себе не является сущностью семантики и его не достаточно для семантики.

Заключение 1. Программы не являются сущностью разума и их наличия не достаточно для существования разума.

СВЕТЯЩАЯСЯ КОМНАТА

Аксиома 1. Электричество и магнетизм — это физические силы.

Аксиома 2. Существенное свойство света — это свечение.

Аксиома 3. Силы сами по себе не являются сущностью эффекта свечения и не достаточны для его наличия.

Заключение 1. Электричество и магнетизм не являются сущностью света и не достаточны для его наличия.



КОЛЕБАНИЯ ЭЛЕКТРОМАГНИТНЫХ СИЛ представляют собой свет, хотя магнит, который перемещает человек, не производит никакого свечения. Аналогично манипулирование

символами в соответствии с определенными правилами может представлять собой разум, хотя у основанной на применении правил системы, находящейся в «китайской комнате» Дж. Сирла, настоящее понимание как будто отсутствует.

Что же мог ответить Максвелл, если бы ему был брошен этот вызов?

Во-первых, он, возможно, стал бы настаивать на том, что эксперимент со «светящейся комнатой» вводит нас в заблуждение относительно свойств видимого света, потому что частота колебаний магнита крайне мала, меньше, чем нужно, приблизительно в 10^{15} раз. На это может последовать нетерпеливый ответ, что частота здесь не играет никакой роли, что комната с колеблющимся магнитом уже содержит все необходимое для проявления эффекта свечения в полном соответствии с теорией самого Максвелла.

В свою очередь Максвелл мог бы «проглотить приманку», заявив совершенно обоснованно, что комната уже полна свечения, но природа и сила этого свечения таковы, что человек не способен его видеть. (Из-за низкой частоты, с которой человек двигает магнитом, длина порождаемых электромагнитных волн слишком велика, а интенсивность слишком мала, чтобы глаз человека мог на них среагировать.) Однако, учитывая уровень понимания этих явлений в рассматриваемый период времени (60-е годы прошлого века), такое объяснение, вероятно, вызвало бы смех и издевательские реплики. «Светящаяся комната! Но позвольте, мистер Максвелл, там же совершенно темно!»

Итак, мы видим, что бедному Максвеллу приходится туго. Все, что он может сделать, это настаивать на следующих трех положениях. Во-первых, аксиома 3 в приведенном выше рассуждении не верна. В самом деле, несмотря на то что интуитивно она выглядит достаточно правдоподобной, по ее поводу у нас невольно возникает вопрос. Во-вторых, эксперимент со светящейся комнатой не демонстрирует нам ничего интересного относительно физической природы света. И в-третьих, чтобы на самом деле решить проблему света и возможности искусственного свечения, нам необходима программа исследований, которая позволит установить, действительно ли при соответствующих условиях поведение электромагнитных волн полностью идентично поведению света. Такой же ответ должен дать классический искусственный интеллект на рассуждение Сирла. Хотя китайская комната Сирла и может показаться «в семантическом смысле темной», у него нет достаточных оснований настаивать, что совершаемое по определенным правилам манипулирование символами никогда не сможет породить семантических явлений, в особенности если учесть, что люди еще плохо информированы и ограничены лишь пониманием на уровне здравого смысла тех семантических и

мыслительных явлений, которые нуждаются в объяснении. Вместо того чтобы воспользоваться пониманием этих вещей, Сирл в своих рассуждениях свободно пользуется отсутствием у людей такого понимания.

Высказав свои критические замечания по поводу рассуждений Сирла, вернемся к вопросу о том, имеет ли программа классического ИИ реальный шанс решить проблему сознательного разума и создать мыслящую машину. Мы считаем, что перспективы здесь не блестящие, однако наше мнение основано на причинах, в корне отличающихся от тех аргументов, которыми пользуется Сирл. Мы основываемся на конкретных неудачах исследовательской программы классического ИИ и на ряде уроков, преподанных нам биологическим мозгом на примере нового класса вычислительных моделей, в которых воплощены некоторые свойства его структуры. Мы уже упоминали о неудачах классического ИИ при решении тех задач, которые быстро и эффективно решаются мозгом. Ученые постепенно приходят к общему мнению о том, что эти неудачи объясняются свойствами функциональной архитектуры МС-машин, которая просто непригодна для решения стоящих перед ней сложных задач.

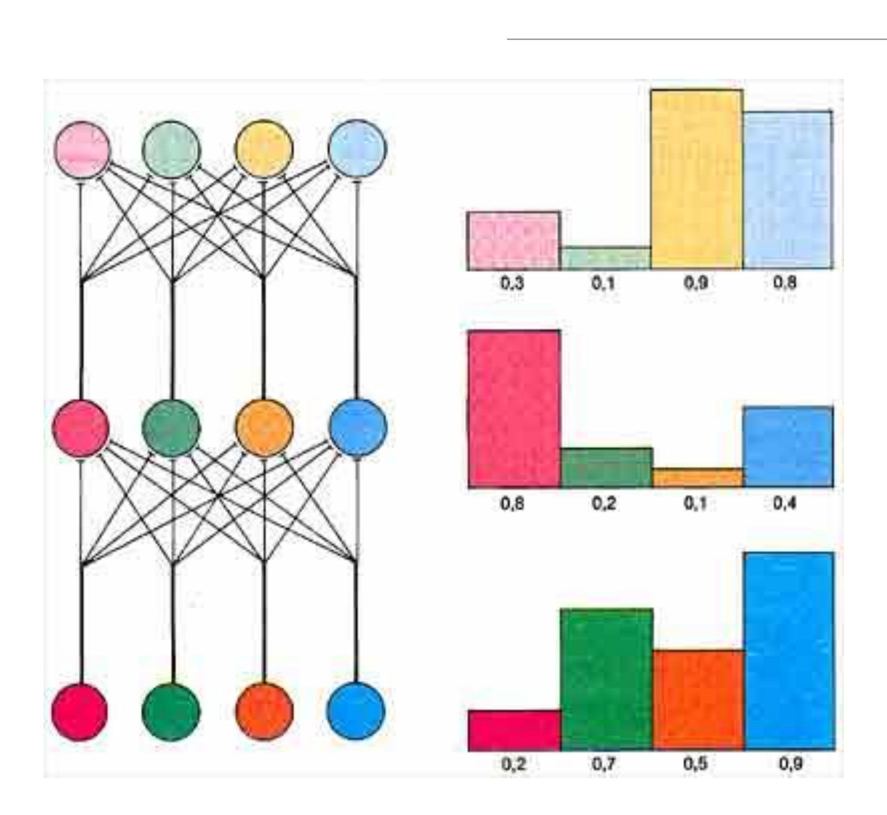
ЧТО НАМ нужно знать, так это каким образом мозг достигает эффекта мышления? Обратное конструирование является широко распространенным приемом в технике. Когда в продажу поступает какое-то новое техническое устройство, конкуренты выясняют, каким образом *оно* работает, разбирая его на части и пытаясь угадать принцип, на котором оно основано. В случае мозга реализация такого подхода оказывается необычайно трудной, поскольку мозг представляет собой самую сложную вещь на планете. Тем не менее нейрофизиологам удалось раскрыть многие свойства мозга на различных структурных уровнях. Три анатомические особенности принципиально отличают его от архитектуры традиционных электронных компьютеров.

Во-первых, нервная система - это параллельная машина, в том смысле, что сигналы обрабатываются одновременно на миллионах различных путей. Например, сетчатка глаза передает сложный входной сигнал мозгу не порциями по 8,16 или 32 элемента, как настольный компьютер, а в виде сигнала, состоящего почти из миллиона отдельных элементов, прибывающих одновременно к окончанию зрительного нерва (наружному коленчатому телу), после чего они также одновременно, в один прием, обрабатываются мозгом. Во-вторых, элементарное «процессорное устройство» мозга, нейрон, отличается относительной простотой. Кроме того, его ответ на входной сигнал - аналоговый, а не цифровой, в том смысле, что частота выходного сигнала изменяется непрерывным образом в зависимости от входных сигналов.

В-третьих, в мозге, кроме аксонов, ведущих от одной группы нейронов к другой, мы часто находим аксоны, ведущие в обратном направлении. Эти возвращающиеся отростки позволяют мозгу модулировать характер обработки сенсорной информации.

Еще важнее то обстоятельство, что благодаря их существованию мозг является подлинно динамической системой, у которой непрерывно поддерживаемое поведение отличается как очень высокой сложностью, так и относительной независимостью от периферийных стимулов. Полезную роль в изучении механизмов работы реальных нейронных сетей и вычислительных свойств параллельных архитектур в значительной мере сыграли упрощенные модели сетей. Рассмотрим, например, трехслойную модель, состоящую из нейроноподобных элементов, имеющих аксоноподобные связи с элементами следующего уровня. Входной стимул достигает порога активации данного входного элемента, который посылает сигнал пропорциональной силы по своему «аксону» к многочисленным «синаптическим» окончаниям элементов скрытого слоя. Общий эффект заключается в том, что та или иная конфигурация активирующих сигналов на множестве входных элементов порождает определенную конфигурацию сигналов на множестве скрытых элементов.

То же самое можно сказать и о выходных элементах. Аналогичным образом конфигурация активирующих сигналов на срезе скрытого слоя приводит к определенной картине активации на срезе выходных элементов. Подводя итог, можно сказать, что рассматриваемая сеть является устройством для преобразования любого большого количества возможных входных векторов (конфигураций активирующих сигналов) в однозначно соответствующий ему выходной вектор. Это устройство предназначено для вычисления специфической функции. То, какую именно функцию оно вычисляет, зависит от глобальной конфигурации синаптической весовой структуры.



НЕЙРОННЫЕ СЕТИ моделируют главное свойство микроструктуры мозга. В этой трехслойной сети входные нейроны (слева *внизу)* обрабатывают конфигурацию активирующих сигналов *{справа внизу)* и передают их по взвешенным связям скрытому слою. Элементы скрытого слоя суммируют свои многочисленные входы, образуя новую конфигурацию сигналов. Она передается внешнему слою, выполняющему дальнейшие преобразования. В целом сеть преобразует любой входной набор сигналов в соответствующий выход, в зависимости от расположения и сравнительной силы связей между нейронами.

Существуют разнообразные процедуры для подбора весов, благодаря которым можно сделать сеть, способную вычислить почти любую функцию (т. е. любое преобразование между векторами). Фактически в сети можно реализовать функцию, которую даже нельзя сформулировать, достаточно лишь дать ей набор примеров, показывающих, какие лары входа и выхода мы хотели бы иметь. Этот процесс, называемый «обучением сети», осуществляется путем последовательного подбора весов, присваиваемых связям, который продолжается до тех пор, пока сеть не начнет выполнять желаемые преобразования с входом, чтобы получить нужный выход.

Хотя эта модель сети чрезвычайно упрощает структуру мозга, она все же иллюстрирует несколько важных аспектов. Во-первых, параллельная архитектура обеспечивает колоссальное преимущество в быстродействии по сравнению с традиционным компьютером, поскольку многочисленные синапсы на каждом уровне выполняют множество мелких вычислительных операций одновременно, вместо того чтобы действовать в очень трудоемком последовательном режиме. Это преимущество становится все более значительным, по мере того как возрастает количество нейронов на каждом уровне. Поразительно, но скорость обработки информации, совершенно не зависит ни от числа элементов, участвующих в процессе на каждом уровне, ни от сложности функции, которую они вычисляют. Каждый уровень может иметь четыре элемента или сотню миллионов; конфигурация синаптических весов может вычислять простые одноразрядные суммы или решать дифференциальные уравнения второго порядка. Это не имеет значения. Время вычислений будет абсолютно одним и тем же.

Во-вторых, параллельный характер системы делает ее нечувствительной к мелким ошибкам и придает ей функциональную устойчивость; потеря нескольких связей, даже заметного их количества, оказывает пренебрежимо малое влияние на общий ход преобразования, выполняемого оставшейся частью сети.

В-третьих, параллельная система запоминает большое количество информации в распределенном виде, при этом обеспечивается доступ к любому фрагменту этой информации за время, измеряющееся несколькими миллисекундами. Информация хранится в виде определенных конфигураций весов отдельных синаптических связей, сформировавшихся в процессе предшествовавшего обучения. Нужная информация «высвобождается» по мере того, как входной вектор проходит через (и преобразуется) эту конфигурацию связей.

Параллельная обработка данных не является идеальным средством для всех видов вычислений. При решении задач с небольшим входным вектором, но требующих многих миллионов быстро повторяющихся рекурсивных вычислений, мозг оказывается совершенно беспомощным, в то время как классические МС-машины демонстрируют свои самые лучшие возможности. Это очень большой и важный класс вычислений, так что классические машины будут всегда нужны и даже необходимы. Однако существует не менее широкий класс вычислений, для которых архитектура мозга представляет собой наилучшее техническое решение. Это главным образом те вычисления, с которыми обычно сталкиваются живые организмы: распознавание контуров хищника в «шумной» среде; мгновенное вспоминание правильной реакции на его пристальный взгляд, способ бегства при его приближении или защиты при его нападении; проведение различий между съедобными и несъедобными вещами, между половыми партнерами и другими животными; выбор поведения в сложной и постоянно изменяющейся физической или социальной среде; и т. д.

Наконец, очень важно заметить, что описанная параллельная система не манипулирует символами в соответствии со структурными правилами. Скорее манипулирование символами является лишь одним из многих других «интеллектуальных» навыков, которым сеть может научиться или не научиться. Управляемое правилами манипулирование символами не является основным способом функционирования сети. Рассуждение Сирла направлено против управляемых правилами МС-машин; системы преобразования векторов того типа, который мы описали, таким образом, выпадают из сферы применимости его аргумента с китайской комнатой, даже если бы он был состоятелен, в чем мы имеем другие, независимые причины сомневаться.

Сирлу известно о параллельных процессорах, но, по его мнению, они будут также лишены реального семантического содержания. Чтобы проиллюстрировать их неизбежную неполноценность в этом отношении, он описывает второй мысленный эксперимент, на сей раз с китайским спортивным залом, который наполнен людьми, организованным в параллельную сеть. Дальнейший ход его рассуждений аналогичен рассуждениям в случае с китайской комнатой.

На наш взгляд, этот второй пример не так удачен и убедителен, как первый. Прежде всего то обстоятельство, что ни один элемент в системе не понимает китайского языка, не играет никакой роли, потому что то же самое справедливо и по отношению к нервной системе человека: ни один нейрон моего мозга не понимает английского языка, хотя мозг как целое понимает. Далее Сирл не упоминает о том, что его модель (по одному человеку на каждый нейрон плюс по быстроногому мальчишке на каждую синаптическую связь) потребовала бы по крайней мере 10^{14} человек, так как человеческий мозг содержит 10^{11} нейронов, каждый из которых имеет в среднем 10^{3} связей. Таким образом, его система потребует населения 10 000 миров, таких как наша Земля. Очевидно, что спортивный зал далеко не в состоянии вместить более или менее адекватную модель.

С другой стороны, если такую систему все же удалось бы собрать, в соответствующих космических масштабах, со всеми точно смоделированными связями, у нас получился бы огромный, медленный, странной конструкции, но все же функционирующий мозг. В этом случае, конечно, естественно ожидать, что при правильном входе он будет мыслить, а не наоборот, что он на это не способен. Нельзя гарантировать, что работа такой системы будет представлять собой настоящее мышление, поскольку теория векторной обработки может неадекватно отражать работу мозга. Но точно так же у нас нет никакой априорной гарантии, что она не будет мыслить. Сирл еще раз ошибочно отождествляет сегодняшние пределы своего собственного (или читательского) воображения с пределами объективной реальности.

МОЗГ является своеобразным компьютером, хотя большинство его свойств пока остается непознанным. Охарактеризовать мозг как компьютер далеко не просто, и такие попытки не следует считать излишней вольностью. Мозг дейтствительно вычисляет функции, но не такие, как в прикладных задачах, решаемых классическим искусственным интеллектом. Когда мы говорим о машине как о компьютере, мы не имеем в виду последовательный цифровой компьютер, который нужно запрограммировать и которому свойственно четкое разделение на программную часть и аппаратную; мы не имеем также в виду, что этот компьютер манипулирует символами или следует определенным правилам. Мозг — это компьютер принципиально другого вида.

Каким образом мозг улавливает смысловое содержание информации, пока не известно, однако ясно, что проблема эта выходит далеко за рамки лингвистики и не ограничивается человеком как видом. Маленькая кучка свежей земли означает, как для человека, так и для кайота, что где-то поблизости находится суслик; эхо с определенными спектральными характеристиками означает для летучей мыши присутствие мотылька. Чтобы разработать теорию формирования смыслового содержания, мы должны больше знать о том, как нейроны кодируют и преобразуют сенсорные сигналы, о

нейронной основе памяти, об обучении и эмоциях и о связи между этими факторами и моторной системой. Основанная на нейрофизиологии теория понимания смысла может потребовать даже наших интуитивных представлений, которые сейчас кажутся нам такими незыблемыми и которыми так свободно пользуется Сирл в своих рассуждениях. Подобные пересмотры — не редкость в истории науки.

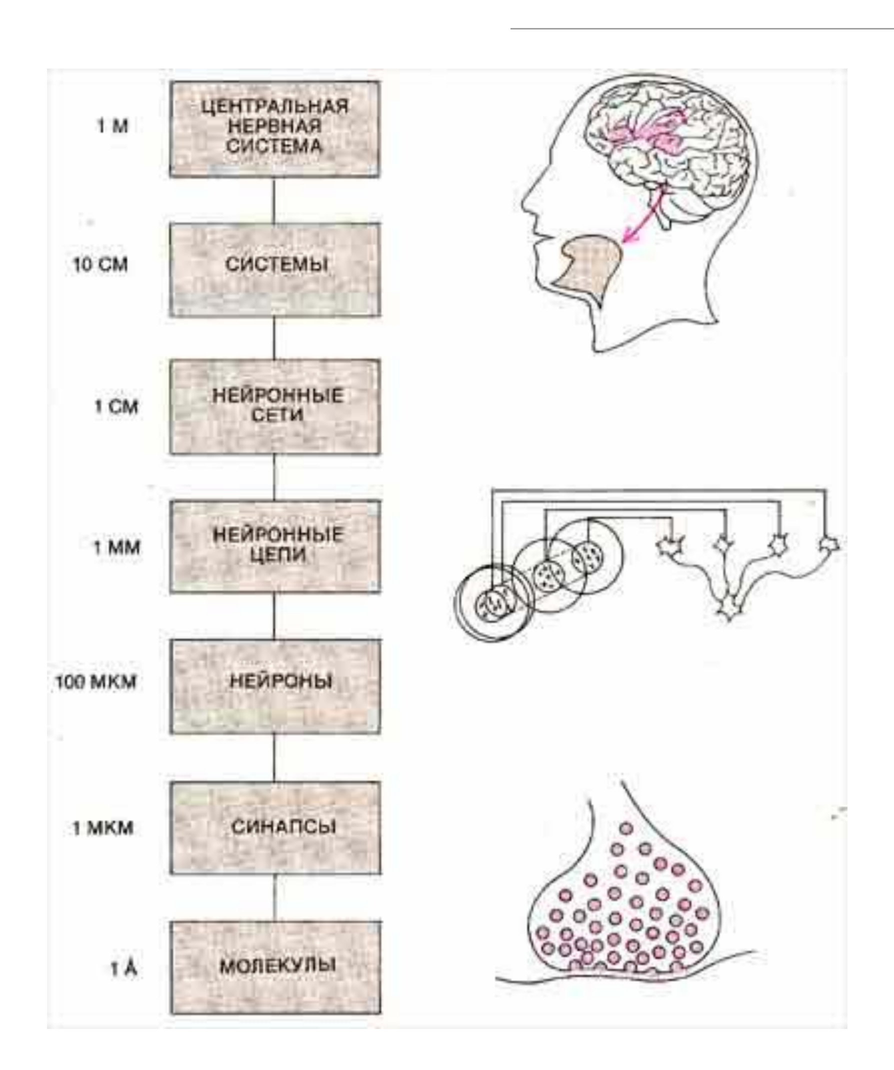
Способна ли наука создать искусственный интеллект, воспользовавшись тем, что известно о нервной системе? Мы не видим на этом пути принципиальных препятствий. Сирл будто бы соглашается, но с оговоркой: «Любая другая система, способная порождать разум, должна обладать каузальными свойствами (по крайне мере), эквивалентными соответствующим свойствам мозга». В завершение статьи мы и рассмотрим это утверждение. Полагаем, что Сирл не утверждает, будто успешная система искусственного интеллекта должна непременно обладать всеми каузальными свойствами мозга, такими как способность чувствовать запах гниющего, способность быть носителем вирусов, способность окрашиваться в желтый цвет под действием пероксидазы хрена обыкновенного и т. д. Требовать полного соответствия будет все равно, что требовать от искусственного летательного аппарата способности нести яйца.

Вероятно, он имел в виду лишь требование, чтобы искусственный разум обладал всеми каузальными свойствами, относящимися, как он выразился, к сознательному разуму. Однако какими именно? И вот мы опять возвращаемся к спору о том, что относится к сознательному разуму, а что не относится. Здесь как раз самое место поспорить, однако истину в данном случае следует выяснять эмпирическим путем — попробовать и посмотреть, что получится. Поскольку нам так мало известно о том, в чем именно состоит процесс мышления и семантика, то всякая уверенность по поводу того, какие свойства здесь существенны, была бы преждевременной. Сирл несколько раз намекает, что каждый уровень, включая биохимический, должен быть представлен в любой машине, претендующей на искусственный интеллект. Очевидно, это слишком сильное требование. Искусственный мозг может и не пользуясь биохимическими механизмами, достичь того же эффекта.

Эта возможность была продемонстрирована в исследованиях К.Мида в Калифорнийском технологическом институте. Мид и его коллеги воспользовались аналоговыми микроэлектронными устройствами для создания искусственной сетчатки и искусственной улитки уха. (У животных сетчатка и улитка не являются просто преобразователями: в обеих системах происходит сложная параллельная обработка данных.) Эти устройства уже не простые модели в миникомпьютере, над которым посмеивается Сирл; они представляют собой реальные элементы обработки информации, реагирующие в реальное время на реальные сигналы: свет — в случае сетчатки и звук — в случае улитки уха. Схемы устройств основаны на известных анатомических и физиологических свойствах

сетчатки кошки и ушной улитки сипухи, и их выход чрезвычайно близок к известным выходам органов, которые они моделируют.

В этих микросхемах не используются никакие нейромедиаторы, следовательно, нейромедиаторы, судя по всему, не являются необходимыми для достижения желаемых результатов. Конечно, мы не можем сказать, что искусственная сетчатка видит что-то, поскольку ее выход не поступает на искусственный таламус или кору мозга и т. д. Возможно ли по программе Мида построить целый искусственный мозг, пока не известно, однако в настоящее время у нас нет свидетельств, что отсутствие в системе биохимических механизмов делает этот подход нереалистичным.



НЕРВНАЯ СИСТЕМА охватывает ыного масштабов организации, от молекул нейромедиаторов (внизу) до всего головного и спинного мозга. На промежуточных у ровнях находятся отдельные нейроны и нейронные цепи, подобные тем, что реализуют избирательность восприятия зрительных стимулов (в *центре),* и системы, состоящие из многих цепей, подобных тем, что обслуживают функции речи (справа вверху). Только путем исследований можно установить, насколько близко искусственная система способна воспроизводить биологические системы, обладающие разумом.

ТАК ЖЕ как и Сирл, мы отвергаем тест Тьюринга как достаточный критерий наличия сознательного разума. На одном уровне основания для этого у нас сходные: мы согласны, что очень важно, каким образом реализуется функция, определенная по входу-выходу; важно, чтобы в машине происходили правильные процессы. На другом уровне мы руководствуемся совершенно иными соображениями. Свою позицию относительно присутствия или отсутствия семантического содержания Сирл основывает на интуитивных представлениях здравого смысла. Наша точка зрения основана на конкретных неудачах классических МС-машин и конкретных достоинствах машин, архитектура которых ближе к устройству мозга. Сопоставление этих различных типов машин показывает, что одни вычислительные стратегии имеют огромное и решающее преимущество над другими в том, что касается типичных задач умственной деятельности. Эти преимущества, установленные эмпирически, не вызывают никаких сомнений. Очевидно, мозг систематически пользуется этими вычислительными преимуществами. Однако он совершенно не обязательно является единственной физической системой, способной ими воспользоваться. Идея создания искусственного интеллекта в небиологической, но существенно параллельной машине остается очень заманчивой и в достаточной мере перспективной.
